University of Gondar

Institute of Public Health

College of Medicine and Health Sciences

Application of Data Mining Techniques to Identify Patterns in Voluntary Counseling and Testing: A Case of Gondar University Hospital VCT Center, Gondar, Northwest Ethiopia

By: Kinfe Wubetu

Advisor(s):-

Dr. Berihun Megabiaw (MD, MPH)

Ato Atinkut Alamirrew (B.Sc., MPH-HI)

A thesis submitted to Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, in partial fulfillment of the requirements for the degree of Master of Public Health in Health Informatics

June, *2012*

Gondar, Ethiopia

UNIVERSITY OF GONDAR

COLLEGE OF MEDICINE AND HEALTH SCIENCES

INSTITUTE OF PUBLIC HEALTH

Title: Application of Data Mining Techniques to Identify Patterns in Voluntary Counseling and Testing: A Case of Gondar University Hospital VCT Center, Gondar, Northwest Ethiopia

By: Kinfe Wubetu

Approved by the Examining Board

_____

Head, Institute of Public Health

Advisors

1. Dr. Berihun Megabiaw (MD, MPH)

2. Ato Atinkut Alamirrew (B.Sc., MPH-HI)

Examiners

1._____

2._____

Dedication

This work is dedicated to my mother, W/ro Yeserash Yeayeneabeba, who gives me love and support in all my day to day life, and my sisters and brothers especially Mulumebet Wubetu for her priceless support throughout the research process.

Acknowledgment

My first deep gratitude and credit goes to my advisors Dr. Berihun Megabiaw (MD, MPH) and Ato Atinkut Alamirrew (B.Sc., MPH-HI) for your advice and unreserved assistance you offered me in finalizing this work. You managed to help me and replay to me even when you were so busy.

I am also grateful to thank Dr. Million Meshesha (PhD) who gave me lecture on data mining course and provided me helpful consultation support throughout this research.

I would also like to thank Gondar University Hospital VCT Center counselors and domain experts for their support in providing the necessary information throughout this research.

I want finally like to pass my heartfelt gratitude to all my colleagues, friends and classmates for their support and encouragement they provided.

Acronyms

AIDS	Acquired Immuno-Deficiency Syndrome
ART	Anti Retroviral Therapy
CART	Classification And Regression Trees
CDC	Center for Disease Control and prevention
CSV	Comma Separated Version
FGA –E	Family Guidance Association -Ethiopia
FMOH	Federal Ministry of Health
HAPCO	HIV/AIDS Prevention and Control Office
HMIS	Health Management Information System
HIV	Human Immunodeficiency Virus
KDD	Knowledge Discovery in Database
NGO	Non-Governmental Organization
PICT	Provider Initiated Counseling and Testing
RPO	Research and Publication Office
STIs	Sexually Transmitted Infections
SVM	Support Vector Machine
UNAIDS	Joint United Nation program on HIV/AIDS
VCT	Voluntary Counseling and Testing
WEKA	Waiikato Environment for Knowledge Analysis

Table of Contents **Page**

List of Tables

List of Figures

List of Annexes **Page**

Abstract

Background: - Voluntary Counseling and Testing (VCT) is an important intervention and entry point in the prevention, control and management of HIV. But the documentation process of VCT logbooks and dataset has given little attention on the identification of the VCT user's and test result patterns in Ethiopia. So, assessment of patterns in VCT outcomes using classification and association techniques of data mining is important for proper intervention.

Objective: - To apply data mining in identifying patterns of VCT dataset to discover knowledge that enables to design proper counseling and prevention strategies.

Method: - The study design was record review, hybrid model of data mining on Gondar University VCT service. The research is based on 12,033 recorded data. Redundant, noisy, misclassified, inconsistent, outlier data cleaned and transformed into the appropriate format used for data mining classification and association rule discovery algorithms by using WEKA software. Missing, misclassified, noisy values were handled appropriately during pre-processing step.

Results

The maximum correctly classified percentage obtained is 85.15% with j482 classification algorithm. And, the optimal attributes of the dataset that has been shown in the experiments to classify HIV test result are physical status of the clients, reasons to test, marital status, history of STIs, age and sex. And as of the domain experts and the researcher, there are some unraveling rules from the classifier. In addition, Apriori association technique experimented with selected variables and positive and negative test results shown various interesting patterns.

Conclusion and Recommendations

In this study, the application of both classifier and association techniques of data mining has shown interesting patterns. Thus, HIV care and support as well as prevention activities shall focus on variables and instances that are optimal and have high support and confidence with positive result and prevention activities shall focus on variables that associated with negative test results.

1. INTRODUCTION

1.1 Statement of the problem

Despite all the efforts to prevent HIV, the pandemic continues to threaten sub-saharan Africa countries including Ethiopia. According to Joint United Nation program on HIV/AIDS (UNAIDS) 2010 report (1), sub-saharan Africa countries bear about 68% of the worldwide HIV/AIDS burden. Estimates suggest that there were approximately 1,296,908 Ethiopians living with HIV. Based on the estimation, the mean national adult HIV prevalence rate for the year 2010 was 2.4% (2).

In Ethiopia, with the understanding of the multifaceted effects of HIV/AIDS, efforts have been made by governmental and non-governmental organizations to avert the adverse effects of the pandemic. The response to HIV/AIDS in Ethiopia has been guided by the national policy issued in 1998.Providing voluntary counseling and testing has been indicated as primary strategy in the policy to mitigate the spread of HIV (3). VCT is the process, by which self initiated individual undergoes counseling that enables him / her to make an informed choice about being tested for HIV. It is an important intervention and entry point in the prevention, control and management of the human immunodeficiency virus (HIV).

VCT service was expanded in the early 1990s by Federal Ministry of Health in collaboration with local and international non-government organizations (4). In December 2009, 1,823 health facilities were provided VCT services in Ethiopia by government and privately owned institutions through integrating it into public health facilities such as hospitals, health centers, and clinics (5).

Though there are an ongoing and courageous works on the availability and accessibility of VCT service, the documentation process of VCT logbooks in the health institutions has given little attention on the identification of the VCT user patterns and the outcome of the test results in Ethiopia. Two of the major challenges mentioned to depict the pattern of users were lack of conceptualization of VCT by program implementers/coordinators and underdeveloped management information systems (6).

In this regard, data mining is a powerful and useful technology with great potential application for the extraction of hidden patterns or knowledge from large data bases thereby organizations can design proper intervention strategies to deal with gaps up on the pattern identified (7).

Classification and association rule discovery techniques of data mining are, therefore, used to assess patterns in VCT outcomes based on the dataset registered at the VCT logbook of the hospital. In case of classification algorithms, the results can identify/classify the attributes and instances that are optimal to positive and negative results that program coordinators and other responsible bodies shall focus on their prevention as well as care and support efforts. On the other hand, in addition to the prevention and care and support program, association technique result can identify attributes and instance that responsible bodies should focus during promotion of VCT services for the future.

1.2 Literature Review

1.2.1 VCT Dataset and Statistical Analysis in Ethiopia

Various studies on VCT in different countries show that considerable proportion of information on the levels and demographic determinants of utilization and outcome of VCT draws from studies of specific groups at risk. Only few analyses have investigated differentials in utilization to derive general patterns, such as the association of education, income, or gender with behavior regarding testing. Most studies tend to focus on measuring the statistical associations among variables that are detached from their social context (8).

Results of studies using the usual regression statistical analysis conducted in health centers at Addis Ababa and 'Burie' in Ethiopia on VCT data showed, sex and employment status had considerable and significant association with positive results with high prevalence in female and unemployed clients (9)(10). Another study also assessed educational level association with HIV prevalence by using dataset from 28 Family Guidance Association of Ethiopia VCT center in Ethiopia (11).According to the study, HIV prevalence decreases significantly with an increase in educational level for both men and women.

1.2.2. Data Mining

Historically, the notion of finding useful patterns in data has been given a variety of names, such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. So that data mining has popular interchangeable usage with knowledge discovery in databases (KDD), but there is a difference between data mining and KDD. While KDD refers to the whole process of changing low level data into high level knowledge, data mining constitutes one of the phases in this process, namely "the application of specific algorithms for extracting patterns from data"(12) .

However, in industry, media and in the database research setting, the term data mining is becoming more popular than the longer term of knowledge discovery from data. Therefore, the researcher in this study has used the generally accepted definition of data mining, which is the set of procedures and techniques for discovering and describing patterns and trends in data (13).

Data mining often follows one of several methods such as predictive modeling, which includes classification and regression techniques that are used to forecast unknown outcomes or variables, and clustering establishes clusters (based on an analogous statistic) in order to compare historical data to analyze new data. In addition, it includes association rule discovery in order to determine relationships between attributes in a database that often can be correlated as a result of causality and Summarization that – generalizes task-specific data into a data cube that can create graphical representations (14).

Hence, data mining is lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. It is concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners (15).

There were top ten data mining algorithms identified by the Institute of Electrical and Electronic Engineering (IEEE) that held on the International Conference on Data Mining (ICDM) in December 2006. Among these are C4.5, *K*-Means, Support Vector Machine (SVM), Apriori, AdaBoost, k-nearest Neighbor (*k*NN(16).

In case of classification algorithm of constructing a decision tree process, first select an attribute to place at the root node and make one branch for each possible value. This splits up the example set into subsets, one for every value of the attribute. The process can be repeated recursively for each branch, using only those instances that actually reach the branch. If at any time all instances at a node have the same classification, stop developing that part of the tree.

The Association rules also are like classification rules that find the rules the same way, by executing a divide-and-conquer rule-induction procedure for each possible

expression that could occur on the right-hand side of the rule. But not only might any attribute occur on the right-hand side with any possible value; a single association rule often predicts the value of more than one attribute. To find such rules, you would have to execute the rule-induction procedure once for every possible combination of attributes, with every possible combination of values, on the right-hand side. That would result in an enormous number of association rules, which would then have to be pruned down on the basis of their coverage (the number of instances that they predict correctly) and their accuracy (the same number expressed as a proportion of the number of instances to which the rule applies)(17).

1.2.3. Data Mining on Public Health Domain including HIV/AIDS Database

Data mining can be applied in various areas of health care management. For example, decision tree classification algorithm on diabetes mellitus type-I database in India with eight attributes/variables could diagnose nephropathy disease early for prevention (18). The same algorithm also applied explaining women's choice of contraceptive methods by using an Indonesian research dataset showed interesting patterns among husband education ,choice of the women's and other attributes like "role power" of motherhood (19).

Similarly, classification data mining rules applied in Korean public healthcare center database to answer research questions of 'can we predict whether a patient revisit a healthcare center of the patients' showed that the classification models can help public healthcare centers plan and implement health care service programs which are more appropriate to the local residents (20).In other study, data mining application for infection control surveillance in hospitals of United State of America using Apriori association rule algorithm expresses an interesting association patterns between species of the organism, drug used, hospital wards, gender, age group and other attributes with low support and confidence (21).Moreover, to improve data collection and information retrieval for monitoring of sexual health, data mining

recommended to be applying on genitourinary medicine clinics in England that have large databases(22).

There are also applications of data mining on HIV/AIDS related databases. Classification and regression trees (CART) on workforce analysis applied on health professionals' population ratio to understand HIV/AIDS prevalence patterns among 194 countries by merging different dataset. By using 19 variables from various global level data repositories, the main factors in understanding HIV/AIDS prevalence rates were identified; these are physician density followed by female literacy rates and nursing density in the countries (23).

Another study that applied data mining analysis on 250,000 anonymised records of AIDS patients' database in Thailand showed that association rule algorithm identified associations that were not expected in the data. It was depicted that the result from the data mining were different from traditional reporting mechanisms utilized by medical practitioners. It also allowed the identification of symptoms that co-exist or are precursors of other symptoms (24).

Prediction of HIV status using Neural Networks classification algorithm was applied on the experimental data obtained from antenatal sero-prevalence surveys conducted in South Africa that contains variables such as age, education, location, race, parity and gravidity. The results showed that HIV status of an individual can be predicted using demographic data to 84.24% accuracy (25).

Although researches with the application of data mining on HIV/AIDS and VCT are hardly available in Ethiopia, there are few studies made on other public health matters. Among these, classification rule of data mining to predict the risk of child mortality obtained by using the decision tree approach provided simple rules that can be used by non technical health care professionals to identify cases for which the rule is applicable (26). The other study on public health issues in Ethiopia was the application of classification decision tree algorithm on road traffic accident showed that the algorithms can about 80% correctly classified the patterns among the severity of the accidents and nine optimal attributes included(27).

There is unpublished research paper that focused on the application of data mining technology on VCT dataset to identify determinants risk factors of HIV infection and to find their association rules, in case of Center for Disease Control (CDC) and prevention in Zewditu Hospital. In this research, some interesting patterns were identified by the health experts and the researcher that helps for prevention of HIV and promotion of VCT services. Among these, variation in HIV infection among the unemployed especially in female was observed. 98.86% of people who previously known their positive sero-status have confirmed exactly. Besides, people whose reason for test was having risk suspect or symptoms is also associated to HIV negative result with promising evidence. Observed that, above 75% of the data was collected from unmarried people that showed unmarried individuals are concerned about HIV case than others and the result showed the probability is only 11.61% which is much smaller than the general probability i.e 17.06%(28).

1.3. Justification

Gondar University Hospital has been providing VCT service since January 2002. Till the month of June 2010, the VCT center used VCT logbook and there were 29,176 registered VCT clients. Although the hospital has organized VCT data, as long as the researcher's knowledge is concerned, there has not been any attempt to characterize patterns of the clients and use for HIV prevention, care and support programs. Consequently, the application of data mining technique to identify different patterns of VCT users is very necessary for prevention and control efforts. This research is then designed to apply data mining techniques in order to bring the shelved VCT clients' dataset to use for proper planning of prevention, care and support by planners, decision makers and program coordinators.

The predominant technique in the analysis of epidemiological studies was linear or logistic regression, in which the effects predictors are linear. It is difficult to use this analysis to discover unanticipated complex relationships. Specifically, as the volume of data increases, the usual statistical method becomes inefficient. This in turn calls

the application of new methods and tools that can help to search large quantities of epidemiological data and to discover new patterns and relationships that are hidden in the data that can be easily understandable. This is why the present researcher focuses on data mining techniques.

2. OBJECTIVES

2.1. General Objective

To apply data mining techniques in identifying patterns of Voluntary Counseling and Testing (VCT) dataset to discover knowledge that enables to design proper counseling and other HIV/AIDS programs.

2.2. Specific objectives

- Generate cleaned and formatted training and testing datasets by applying data preprocessing and transformation techniques
- Identify different patterns of VCT client attributes by using classification and association rule discovery techniques
- To extract knowledge that support the VCT services and other HIV/AIDS programs

3. METHODS

3.1. Study Design

The study design of this research was institution based record review on VCT service dataset with hybrid data mining model.

3.2. Study Setting and Area

The study was conducted at Gondar University Hospital, Voluntary Counseling and Testing Center. The hospital is located in Gondar town 730 kilometer away from Addis Ababa. The VCT center has been providing VCT services since January 2002. Currently it is providing VCT service for about 25-30 clients per day.

3.3. Source and Study Population/Dataset

This research was based on recorded data available at Gondar University Hospital VCT center that are registered on the VCT logbooks from September 2007 to June 2010. The total number of VCT users in this period were12, 033 and it was considered for this study.

3.3.1. Inclusion Criteria

Based on the discussion with domain experts, dataset with instance or row that has more than 75% of the attributes or columns completed (at least eight attributes).

3.3.2. Exclusion criteria

Data with instances or rows that has no HIV test result recorded or unreadable.

3.4. Sampling Method

Since there was no significant change of association and classifier rules and parameters value after using the three years dataset, 12,033 entered to excel form from 29,176 registered VCT clients from January 2002 –June 2010 and used. This data was used to create the model with the help of various data mining algorithms integrated in WEKA (Waikato Environment for Knowledge Analysis) Software (17). Since the researcher is familiar with this open source software which is a popular suite of machine learning software that is designed for mining data and accordingly uncovers knowledge from huge datasets.

3.5. Variables of the study/Attributes

Dependant variables/classes

- HIV sero status (HIV+ & HIV-)

Independent variables

- Socio-demographic (Age, Sex, Marital Status, Educational Status, Residence(woreda/town) and Religion)
- Economic characteristics (Occupation and income/month for permanent employee)
- Members (No) of households
- Present health status at the time of testing(sick/ok)
- History of STI(Yes/No)
- Reasons for visiting /testing (to know, to confirm, pre-marriage, Reunion, ART)

3.6. Operational Definition

VCT Pattern is an order or arrangement of voluntary counseling and testing attributes (value of variables) in the association or classification rules discovery algorithms.

Data mining is the set of procedures and techniques for discovering and describing patterns and trends in data.

Physical status ok is when the client visited the VCT center he/she was ambulatory and there was no other illness report from him/her.

3.7. Data collection procedure and quality control

3.7.1. Data Collection Procedure

VCT client included in the present research were those actually registered in a standard VCT logbook. This recorded dataset was copied to the prepared data collection format based on the VCT logbook that excludes the name of tested individuals. Two individuals who are working in the hospital as peer counselors have been oriented for half day by the investigator and counselor at the hospital on how to collect the data from the logbook. Thus, the whole data collection was supervised by a senior counselor.

3.7.2. Data Management

Two data clerks were selected on the basis of their experience on data entry of such large dataset and clerks hired for this task took one day training on how to enter the data and keep the data entry quality.

The data clerks had day to day telephone and in person communication with the investigator during the whole data entry process and during difficulties to enter data. Besides, the investigator cross checked data entered everyday for the first five working days and kept doing it on every weekend on randomly selected dataset to ensure quality of the data entry.

Entry of the dataset made by using Microsoft® excel format and was converted to Microsoft ®Comma separated version (CSV) and then to attribute-relation file format (.arff) for the appropriate use of WEKA (Excel → CSV → .arff).

3.8. Data processing and analysis

The patterns discovered from this study are expected to be actionable to solve domain-specific problems, and taken as grounds for performing effective actions. To make domain-driven data mining effective, user guides and intelligent human-machine interaction interfaces were essential through incorporating both human qualitative intelligence and machine quantitative intelligence.

As a reason, hybrid model of data mining processing was applied. Thus, based on the suggested hybrid methodology of data mining process, to this specific data-mining research, the following steps were undertaken (Annex I -Diagrammatic representation of Hybrid Data Mining Process):

Step I- Understanding of the problem domain.

In this step, the researcher made consultation with the counselors, responsible persons for planning and implementing HIV/AIDS programs at North Gondar Zone health Department and Family Guidance Association, HIV/AIDS experts (non-governmental organization) on the problem domain (HIV/AIDS) before starting the data mining process (Annex II-List of domain experts). Unstructured questionnaire was prepared and interviews were conducted on the attributes of the dataset and previous use of VCT patterns in their respective organization, if there was. The responses of some domain experts were summarized and used for the data mining process.

Most of the domain experts mentioned that they have been using the VCT data only for fulfilling the reporting requirements for higher bodies and they specifically tried to categorize the VCT result based on age and sex groups. Though they couldn't analyze the patterns of VCT users and test results, they agree on the need of information based prevention and care and support strategies, So that identifying the

VCT users and test result patterns can be an important input in future planning and implementation of prevention, promotion as well as care support of the organizations program.

Step II- Understanding of the data

In addition to checking the data quality and completeness manually, after the data entry redundancy, missing values were checked by using Explorer of WEKA Software. Accordingly, missing values of nominal variables such as gender, residence, reason to test filled with the group mode of the attributes and the mean age was filled for the missing values of age and missing values for education, occupation and marital status filled based on the age group mode. Other necessary action was taken in order to generate clean data which was useful for the data mining.

Step III- Preparation of the data

This step was so tedious and time consuming in this research process. Noisy, misclassified, inconsistent, outlier data checked and cleaned like occupation of housewife with male sex in the same row , age more than 100 , misclassification of an instance in one column/variable in to other variable and others . Then the cleaned data was further processed by discretization of the age of the clients (to derive new attributes). Hence, the researcher applied the data mining techniques to fulfill the data mining objectives as stated above. For the purpose of applying Apriori association technique, the age of the clients was re-categorized and changed to words based on World Health Organization's (WHO's)classification of five year groups categories.

Step IV- Data mining techniques

The data mining task was done with WEKA software version 3.4. Because, WEKA is open source software with a collection of machine learning algorithms for data mining tasks. In addition, the researcher has sufficient familiarity with WEKA 3.4 version software.

The researcher employed classification/prediction algorithm of the j48, DecisionStump, Lazy.LWL.DecisionStump-W0 and REPTree-M2-v0.0010-N3-S1-D-1 of explorer option and compared the parameters values, since the dataset contain both nominal and numerical values/instances and appropriate for these classification algorithms. In addition association rule discovery algorithm (Apiori) of the WEKA 3.4 version software data mining technique used, since the dataset was appropriate for the application of these algorithms and they are mostly used in the data mining technique (16). The derived knowledge from preprocessed data using the selected data mining techniques was compared by the percentage of correctly classified, mean error, mean square error , confidence and support percentage and others parameters.

Besides to identify the patterns in the whole dataset, the patterns of the dataset by using association rule discovery technique was assessed by dividing positive and negative results of the VCT.

Since the data mining process was iterative; selection of attributes for the pattern identification and subjective interestingness of the patterns , value of support and confidence was guided and assessed by some of the domain experts identified in step one and the researcher.

Step V- Evaluation of the discovered knowledge (patterns)

The evaluation of this study include; understanding of the results, checking whether the discovered knowledge was interesting or not, interpretation of the results by the researcher and domain experts. Discussion was conducted and responses were collected from some of domain experts on the patterns or the discovered knowledge interestingness, interpretation of the results and checked the impact of the discovered knowledge. So that, both subjective and objective measure of rules/patterns interestingness parameters used in the research process.

Step VI- Use of the discovered knowledge.

This final step of data mining consisted of planning where and how to use the discovered knowledge. The application of the data mining results in this specific area will extended to the above mentioned domain experts mentioned in the first step of the data mining process and other relevant bodies that can use the study results.

3.9. Ethical consideration

The ways in which data mining used in the case of this research kept the confidentiality of the data. In addition, the result of the analysis showed rare socio-demographic attributes of the outliers identified and transformed to other category.

As a result, by considering ethical issues, ethical clearance was secured from Institutional Review Board (IRB) of Institute of Public Health, University of Gondar. Permission to proceed and accomplished the study was obtained from the Medical Director, VCT center coordinator and other responsible bodies.

Objectives of the study clearly explained for the responsible bodies to coordinate the Hospital VCT center and it had been properly understood. Confidentiality was maintained by omitted the name of the tested clients when we copied the VCT logbook. Therefore, this study will never cause and harm individuals who were registered as VCT clients and included in the data mining process.

4. RESULTS

Based on the application of the data mining techniques, a total of 13 attributes and 12,033 rows (instances) were entered for analysis.

The experiment starts by understanding the datasets using -WEKA, especially the number of instances of each attribute. Among the 12,033 instances/rows of the dataset, 6,183 (51.2%) were males and the mean age of the clients was 26± 10 SD years.

There were 340 (2.8%) clients who reported previous history of STIs and the number of positive result of the test was 2,093 (17.39%) (See Table 1 and Annex IV).

Table 1: The top two/three percentage of instances of the attributes in the dataset

Attributes / Variables	Instances	Number of Instances	%of Instances
Sex /gender of the client	Male	6183	51.38
	Female	5850	48.62
Residence	Gondar	6955	57.8
	Dembia	789	6.56
	Chilga	686	5.70
Religion	Orthodox	11337	94.22
	Muslim	541	4.50
	Protestant	72	0.60
Marital Status	Single/Never married	6871	57.10
	Married	2671	22.20
	Divorced	1642	13.65
Educational Status	Secondary school	3925	32.62
	Full Primary (grade 5-8)	2322	19.30
	Illiterate	1972	16.39
Occupation/Job	Student	3252	27.03
	Farmer	1709	14.20
	House wife	1564	13.0
History of STIs	No	11693	97.17
	Yes	340	2.83
Physical Status	Ok	11379	94.56
	Sick	654	5.44
Reason to test	To know own status	10411	86.52
	Pre-Marriage screening	1467	12.19
	Partners pressure	66	0.55
Test Result	Negative	9668	80.35
	Positive	2093	17.39
	No result stated/unclear	272	2.26

Under data preprocessing, age was discritized for application of classifier technique and re-categorized and renamed into nominal attribute for association techniques. In addition instances with no result stated or unclear to read excluded before the application of data mining algorithms.

Based on the discussion with domain experts, since 10,663 instances from income per month and 9,413 instances from number of household members were incomplete and they are numerical, the researcher excluded these two attributes for the purpose of association technique.

Classifier Technique

Since the data contains both numerical and nominal data/instances, in using classification package of explorer window, after discritizing age, the researcher applied and compared the j48, DecisionStump, Lazy.LWL.DecisionStump-W0 and REPTree-M2-v0.0010-N3-S1-D-1algorithms of explorer option of classifier package of weka.3-4 version software (See Table 2).

Table 2: Parameters results of the various classifier algorithms using the default test mode

Classification Algorithm with 66 %training dataset	Correctly Classified instance %	Incorrectly Classified instance %	Kappa Statistics	Mean Absolute Error	Root Mean squared error
j48-C0.25-M2	84.99	15.1	0.243	0.258	0.353
DecisionStump	84.77	15.23	0.232	0.265	0.356
Lazy.LWL.DecisionStump-W0	84.77	15.23	0.233	0.260	0.349
REPTree-M2-v0.0010-N3-S1-D-1	84.02	15.98	0.197	0.241	0.359

For the first trial, the researcher used the default test mode of split i.e. 66% train and remaining test for the whole attributes and instances and then with various percentage, resample and AdaBoost, Bagging and other experiments. The

maximum correctly classified percentage obtained at split 55% train, remaining test of j48 model.

Table 3: Summary Table for Analysis Parameters results of the j48 model at 55% and 66% Training and Remaining Test

Parameter	**Percentage of Training and Remaining Test**	
	66% Training and Remaining Test	**55 % Training and Remaining Test**
Correctly Classified Instances	84.99%	85.15%
Kappa Statistics	0.21	0.23
Mean Absolute Error	0.25	0.26
Relative Absolute Error	88.4%	88.1%
Precision for Negative	0.86	0.86
Precision for Positive	0.69	0.71
F-Measure for Negative	0.92	0.92
F-Measure for Positive	0.26	0.28

The result showed with the following model characteristics and 20 rules (See fig.1 and Annex V):

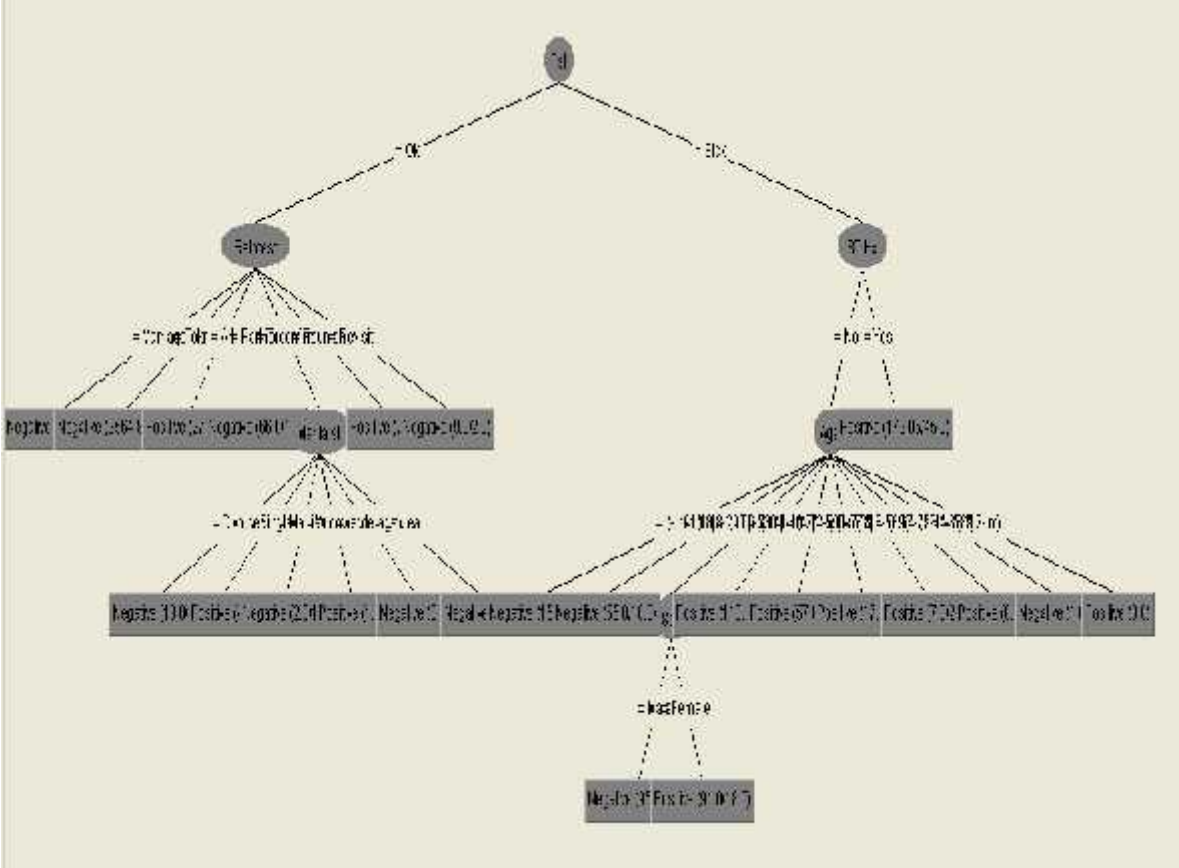


Fig.1: J48 Decision Tree classifier results with the whole dataset

The optimal attributes of the dataset that has been showed in the experiments, to classify/ predict the result of the test, are that of six attributes i.ePst (physical status of the clients), Retotest (reason to test), Maritalst (marital status), STIHx (history of STIs), Age and Sex/gender.

As of the domain experts and the researcher, the unraveling rules results of the classifier that identified as not anticipated and common based on their expectation are:

Rule Number 4: If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to partner then the result is negative (66/1)

Rule Number5: If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to confirm and Maritalst (marital status) is equal to divorced then the result is negative (13/2)

Rule Number6: If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to confirm and Maritalst (marital status) is equal to single then the result is positive (4/1)

Rule Number14: If the Pst (physical status of the clients) is equal to sick and STIHx (history of STIs) is equal to No and age 21-30 and sex is equal to Male then the result is negative (95/12)

Rule Number15: If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 21-30 and sex is equal to Female then the result is positive (91/18)

Rule Number16: If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 30-40 then the result is positive (110/34)

Rule Number17: If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 40-50 then the result is positive (57/25)

Rule Number18: If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 50-60 then the result is positive (17/6)

The remaining rules identified as reveling or expected by the domain experts and the researcher are like:

Rule Number1: If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to marriage then the result is negative (1445/54)

Rule Number2: If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to Tokn (to know) then the result is negative (9525/1604)

Rule Number11: If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to yes then the result is positive (179/45)

Even the accuracy of the j48 model classifier after trying with different percentage values of the test mode of split or folds for the whole dataset, the above model characteristics were almost similar and the maximum results are at split 55 % train.

The following model characteristic observed while using dataset at split 55% train and remaining test of j48 algorithm:

=== Summary ===

Correctly Classified Instances	4507	85.1502 %
Incorrectly Classified Instances	786	14.8498 %
Kappa statistic	0.2298	
Mean absolute error	0.2567	
Root mean squared error	0.3529	
Relative absolute error	88.0916 %	
Root relative squared error	94.7948 %	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.986	0.825	0.858	0.986	0.917	Negative
0.175	0.014	0.708	0.175	0.28	Positive

=== Confusion Matrix ===

a	b	<-- classified as
4354	63	**a = Negative**
723	153	**b = Positive**

Fig.2 Evaluation on split using 55% train, remaining test with j48 classification algorithm

The highest parameters of the classifier obtained on test split at 55% with 85.15% correctly classified instances ,Kappa statistic is 0.229,true positive for negative result is 98.6% and precision for negative class and positive class are 0.858 and 0.708 respectively(See Fig. 2).

But, after discussion held with some domain experts like senior VCT counselor in Gondar University Hospital VCT center and based on the objective of the research, Residence (place of residents), Educst(educational status), and Occupation attributes, after best attribute selection, were included and the application of the classifier model and the experiments applied attribute by attribute for each of them and in combination. And, the result of the classifier by using a separate combination of the attributes which was included by the suggestion of the domain experts was similar with rules by using all attributes in the prediction/ classifier first experiment.

Discovering Association Rules

As association mining with Apriori algorithm by WEKA data mining tool generate best relationship with nominal attributes ,the age of the clients was re-categorized in to five years gap and above sixty five years for those clients who are sixty five and above(see annex VI).

After pre-processing of the dataset, the association mining technique with Apriori algorithm applied for the whole dataset with the parameters that begins with maximum support of 100% of the dataset and decreases this in steps of 5% (default Delta) until there are at least ten rules (by default) with the required minimum confidence is set to 0.9 (by default).

The best five rules that are generated using the Apriori association rule discovery algorithm are stated in Figure 3:

```
=== Associator model (full training set) ===

=== Run information ===

Scheme:  weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation:  VCTTest-weka.filters.unsupervised.attribute.Discretize-B10-Rfirst-last

Instances:   11761

Attributes:  11

Minimum support: 0.85 and Minimum metric <confidence>: 0.9

Best rules found:

 1. Religon=Orthodox Pst=Ok  ==>STIHx=No     conf:(0.99)

 2. Pst=Ok  ==>STIHx=No     conf:(0.99)

 3. Religon=Orthodox  ==>STIHx=No     conf:(0.97)

 4. Religon=Orthodox STIHx=No  ==>Pst=Ok     conf:(0.96)

 5. STIHx=No  ==>Pst=Ok     conf:(0.96)
```

Fig.3TheApriori association of 5best rules for using the whole dataset and11 attributes

The association mining output of Apriori shown in Figure 3. The number preceding ==>symbol indicates the rule's support, i.e the number of items covered by its premise. Following the rule is the number of those items for which the rule's consequently holds as well. In parentheses is the confidence of the rule's that is the second number divided by the first number.

From the first experiment(Figure 3), among the rules found , it was seen (Rule Number1) that if someone is orthodox and his/her physical status is ok, there is 99% confidence with 91.17 % support that the person's STIHx(history of STIs) response is no while the client was visiting the VCT center.

In addition, in consultation with the domain experts and based on the objective of the research, association technique applied for the whole dataset with age, sex,

residence, religion, marital status (Maritalst), education (Educst), occupation/Job, reason to test (Retotest) and result of the test (Result)(See Annex VII). Among the best ten rules; reason to test is to know with negative result, sex male, marital status single and residence in Gondar has 67.9 %, 51.4%, 57.1% and 57.8% support respectively with 94% confidence that the religion of the clients is orthodox.

The next experiments with the association technique were by separating the positive and negative test results.

The best ten association rules of positive test results with Minimum support: 0.8 and Minimum metric <confidence>: 0.9are:

Best ten rules found:

1. Retotest=Tokn 1984 ==> Result=Positive 1984 conf:(1)
2. Religon=Orthodox 1965 ==> Result=Positive 1965 conf:(1)
3. STIHx=No 1883 ==> Result=Positive 1883 conf:(1)
4. Religon=Orthodox Retotest=Tokn 1864 ==> Result=Positive 1864 conf:(1)
5. STIHx=No Retotest=Tokn 1785 ==> Result=Positive 1785 conf:(1)
6. Religon=Orthodox STIHx=No 1770 ==> Result=Positive 1770 conf:(1)
7. Pst=Ok 1693 ==> Result=Positive 1693 conf:(1)
8. Religon=Orthodox STIHx=No Retotest=Tokn 1679 ==> Result=Positive 1679 conf:(1)
9. Religon=Orthodox 1965 ==>Retotest=Tokn Result=Positive 1864 conf:(0.95)
10. Religon=Orthodox Result=Positive 1965 ==>Retotest=Tokn 1864 conf:(0.95)

Fig.4: The Apriori association ten best rules by using the positive result and the 11 attributes

In addition with the consultation with the domain experts and based on the objective of the research association technique experimented with age, sex, marital status (Maritalst), education(Educst), occupation/Job and positive test results (Result)(See

Annex VIII).From the best ten rules that showed 100% confidence with positive result and higher support from the respective clients socio-demographic characteristics are female, married, illiterate, age 25 -29 years and house wife with 55.7 %, 35.3%,29.2% , 23.1% and 20.5% respectively.

In the same manner, the next experiments with the association technique were using negative test results with the selected eleven attributes (See Annex IX). Among the best ten rules that showed 100 % confidence and higher support of the attributes are no STIs history with 98.7% support, ok physical status with 97.6% support, orthodox religion with94.3 % support and reason to know with 84.5% support.

Besides, with the consultation of the domain experts and based on the objective of the research, the association technique on the negative results experimented with age, sex, religion, marital status (Maritalst), education (Educst), occupation/Job and negative test results (Result)(See Annex X).Among the best twenty rules that showed 100% confidence and higher support with the negative result from the respective clients are orthodox religion with 94.3%, single with 62.6% , male with 52.9% , age of 20 -25 years31.7 % , and students with 30.9 % support.

However, among the best ten rules, there are rules with 100% confidence and higher support that showed the occurrence of more than one socio-demographic attributes/ instances in the negative association like orthodox-single with 58.9%support, male –orthodox with 49.9% support and male-single with 34.1% support.

5. DISCUSSION

Most studies on VCT in different countries have focused on specific groups at risk and measured statistical association among variables that are detached from their social context with test outcome(8). In this research, the whole variables that contain socio-demographic characteristics, history of STIs, physical status of the clients while visiting the center and reason for test of dataset are used for data mining tasks.

The WEKA explorer showed that most of the clients for the VCT center were male, age group from 15 to 34 years, Orthodox Christians, who has secondary education and single in marital status. The positive prevalence of the dataset is 17.39% which is almost similar to the case of the study conducted on Zewditu Hospital, VCT center i.e. 17.06% (28).

The highest percentage of accuracy of the classification i.e 85.15% is observed with j48-C0.25-M2 classification algorithm at 55% split that is equivalent to the prediction of HIV status from demographic data using Neural Networks classification algorithm applied on data from antenatal sero-prevalence survey in South Africa i.e. 84.24%(25). And, in case of classifier /prediction technique, physical status of the client, reason to test, marital status, history of sexual transmitted infection (STIs), age and sex are the optimal predictors for the outcome of VCT HIV test.

The first optimal attribute for the classification techniques that showed unrevealing and reveling rules depend on the physical status of the clients while visiting the VCT center. The physical status of the clients as the first classifier attribute indicated that physically ill/sick clients may suspect their sero -status while their health condition may deteriorate and want voluntary HIV counseling and testing. This can also indicate Provider Initiated Counseling and Testing (PICT) for HIV shall strengthen side to side.

Reason to test as classifier/ predication attribute for physically well clients and the client reason with to confirm (Annex V :Rules Number 5 -10) showed that their results depend on the marital status of the client that of divorced, married and under-

aged children have negative test outcome and that of single and widowed marital status showed positive result. Those clients with other reasons except for ART and reunion have negative results.

Rule Number16 If the Pst (physical status of the clients) is equal to sick and STIHx (history of STIs) is equal to No and age 30-40 then the result is positive (110/34) and Rule Number17 If the Pst (physical status of the clients) is equal to sick and STIHx (history of STIs) is equal to No and age 40-50 then the result is positive (57/25) (Annex V), the prediction of HIV test outcome with sick physical status are depend on absence of history of STIs and sex of the clients. According to this study, the prediction of HIV test outcome are positive for females and negative for males which implied females may be more susceptible for HIV and even in the absence of history of STIs. In addition, though the majority of the clients are males, females showed that 100% confidence and 55.7% support that implied and supported the classifier technique that females are more exposed for HIV infection. This is similar to a study in Addis Ababa, Ethiopia (28) which reported variations in HIV infection among the unemployed especially, in female was observed.

In this study history of STIs of the client is one of classifier of the test results and also observed in Apriori association technique in the best rules. Besides, study on genitourinary medicine clinic in England (22) reflected that data mining techniques were applied to improve data collection and information retrieval for monitoring of sexual health. In addition, in generating association rules using Apriori algorithm on the whole dataset, most of the rules showed that absence of history of STIs of the clients has 99% confidence and more than 50% support with single, orthodox, physically well people, the reason of the client to HIV test is to know own status and negative test results that can show most of the VCT clients history of STIs response is no/absence. Thus, it will be of great help if data mining techniques can apply in treated and registered STIs cases in Gondar University Hospital.

Though the instances of age group of 20 -24 are more than age group of 25-29 , but age group of 25-29 has 23% support than 16.6% support of age group of 20 -24 with positive result that showed age group of 25-29 are more exposed to HIV

infection. From the same experiment with 100% confidence, illiterates(29.15%)has more support than secondary school educational status (25.18%) and age group of 20 -24 and female has 11% support that showed there is relatively high prevalence of the infection in these groups(See Annex VII).

Male with married and single marital status with 100% confidence and 17.6 % and 16.7% support respectively showed us most in the dataset both married and single male clients has positive test result outcome (Annex VIII : Rule Number14 and 15). Besides, the association rules of positive test result showed that male and single marital status that has secondary school educational status has 100% confidence with the support of 10.4% and 10.2% respectively (Annex VIII: Rule Number31and 33).

As of the study limitation, most of the instances from income per month and number of household members were not recorded and there were also missing instances in other attributes that shows poor recording system of the VCT center.

6. CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

- In this research, the application of both classifier and association techniques of data mining has showed some interesting patterns that can help planning prevention and care and support of HIV and AIDS programs
- Physical status, reason to test, marital status, history of STIs, age and sex are the optimal attributes for prediction/classifiers of HIV test outcome.
- The majority clients' reason to HIV test is to know that own status is 100% confidence and high support with negative result that showed encouraging attitude of the clients in the dataset.
- From the dataset, most of the clients are males, but the test results has high confidence and support percentage association rule with female and positive test result shows females are more exposed for HIV infection. In the same manner as of the association rule showed, there is high HIV prevalence in25-29 age group.
- There were missing instances in all attributes and in particular on income and family size and in the test results that can shows there is limitation in case of the data recording and management quality of the VCT center.

6.2. Recommendations

HIV /AIDS Program coordinators, VCT centers and Decision makers

- HIV care and support as well as prevention activities shall focus on attributes that are optimal and have high support and confidence with positive result and prevention activities shall focus on attributes associated with negative test results.

Gondar University Hospital

- As per the findings of this study and literature reflected, it will be of great help to further apply data mining techniques on the STIs dataset in the Hospital to improve data collection and information retrieval for monitoring purpose.
- Gondar University Hospital VCT center has to be given due attention to improve data recording system to improve data quality and reducing missing instances and recording all information required on the VCT logbook.

Researchers

- Researchers shall try to apply other classification and association data mining techniques and algorithms on the same dataset to come up with other identified patterns and models taking this study as base.
- As there are interesting patterns in this research paper, the application of data mining techniques shall also be considered in other public health issues including HIV and AIDS related.

7. REFERENCES

1. The Joint United Nations Program on HIV/AIDS (UNAIDS). UNAIDS Report on the Global AIDS Epidemic. Geneva: 2010.

2. Ministry of Health (MoH) and Ethiopia HIV/AIDS Prevention & Control Office (HAPCO). Single point HIV Prevalence Estimate,. Addis Ababa Ethiopia: 2007.

3. Federal Democratic Republic of Ethiopia. Policy on HIV/AIDS of the Federal Democratic Republic of Ethiopia. Aug 1998.

4. National HIV/AIDS Council Secretariat (NACS). National Guideline for VCT in Ethiopia. Oct 2000.

5. Federal HIV/AIDS Prevention and Control Office. Report on progress towards implementation of the UN Declaration of Commitment on HIV/AIDS 2010. March 2010.

6. Ethiopia HIV/AIDS Prevention and Control Office (HAPCO) and Global AIDS Monitoring & Evaluation Team (GAMET). HIV / AIDS in Ethiopia -an Epidemiological synthesis: World Bank Global HIV/AIDS Program. April 2008.

7. Larose. D. T. Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, New Jersey: John Wiley & Sons Inc; 2005.

8. Carla M. O. and Michelle O. The Utilization of Testing and Counseling for HIV: A Review of the Social and Behavioral Evidence. American Journal of Public Health. Oct 2007; 97(10): 1-13.

9. Korra A. Bejiga M. and Tesfaye S. Socio-demographic profile and prevalence of HIV infection among VCT clients in Addis Ababa. Ethiopian Journal of Health and Development. 2005; 19(2):109-116.

10. Biadglegne F, Belyhun Y. and Tessema B. Sero-prevalence of human immunodeficiency virus (HIV) among voluntary counseling and testing (VCT) clients

in Burie Health Center, West Gojjam, Ethiopia. Ethiopian Medical Journal. April 2010; 48(2):149-156.

11. Bradley H., Bedada A, Brahmbhatt H , Kidanu A , Gillespie D , Tsui A Educational attainment and HIV status among Ethiopian voluntary counseling and testing clients. Epub. 2006 Nov 2;11(5):736-42.

12. Fayyad U., Piatetsky-Shapiro G. and Smyth P. From Data Mining To Knowledge Discovery In Databases. American Association for Artificial Intelligence. 1996:37-54.

13. Witten I. H. and Frank E. Data Mining: Concepts and Techniques. Morgan Kaufmann; 2006, 2nd ed.

14. Payton C.F. Data Mining in HealthCare Applications. In: Data Mining: Opportunities and Challenges. Montclair State University, USA: Idea Group; 2003.

15. Hand J.D. Data Mining: Statistics and More? The American Statistician. May 1998 ;52 (2) :114-118.

16. Xindong W. Vipin K., J. Ross Q., Joydeep G., Qiang Y., Hiroshi M., et.al Top 10 algorithms in data mining. Knowl Inf Syst. 2008; 14:1–37.

17. Witten I. H. and Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.

18. Kaur H. and Krishan S.W. Empirical Study on Applications of Data Mining Techniques in Healthcare. Journal of Computer Science. 2006;2 (2):194-200.

19. Bach M.P. and Ćosić D. Data mining usage in health care management: literature surveyand decision tree application. Med glas. 2008;5(1):57-64.

20. Choi K., Chung S., Rhee H. and Suh Y. Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers. The Korean Society of Medical Informatics. June 2010;16 (2):67-76.

21. Ma L. A, Tsui F., Hogan W.R., and Wagner M.M. ,Framework for Infection Control Surveillance Using Association Rules. 2000; Available on http://rods.health.pitt.edu/LIBRARY/Ma%20Lili%20AMIA%202003paper9.pdf Accessed on March 2011

22. Lau R.K.W. and Catchpole M. Improving data collection and informationretrieval for monitoring sexual health. International Journal of STD & AIDS. 2001;12:8-13.

23. Elizabeth A M. , Olivier L. C. and Miklos Z. Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns. BioMed Central Ltd. 2008.

24. Vararuk A., Petrounias I and Kodogiannis V. Data mining techniques for HIV/AIDS data management in Thailand. Journal of Enterprise Information Management. 2008;21 (1):52-70.

25. Brain L .B,Tshilidzi M., Taryn T, Monica L. Prediction of HIV Status from Demographic Data Using NeuralNetworks, In: Proceedings2006 IEEE International Conference onSystems, Man, and CyberneticsOctober 8-11, 2006, Taipei, Taiwan

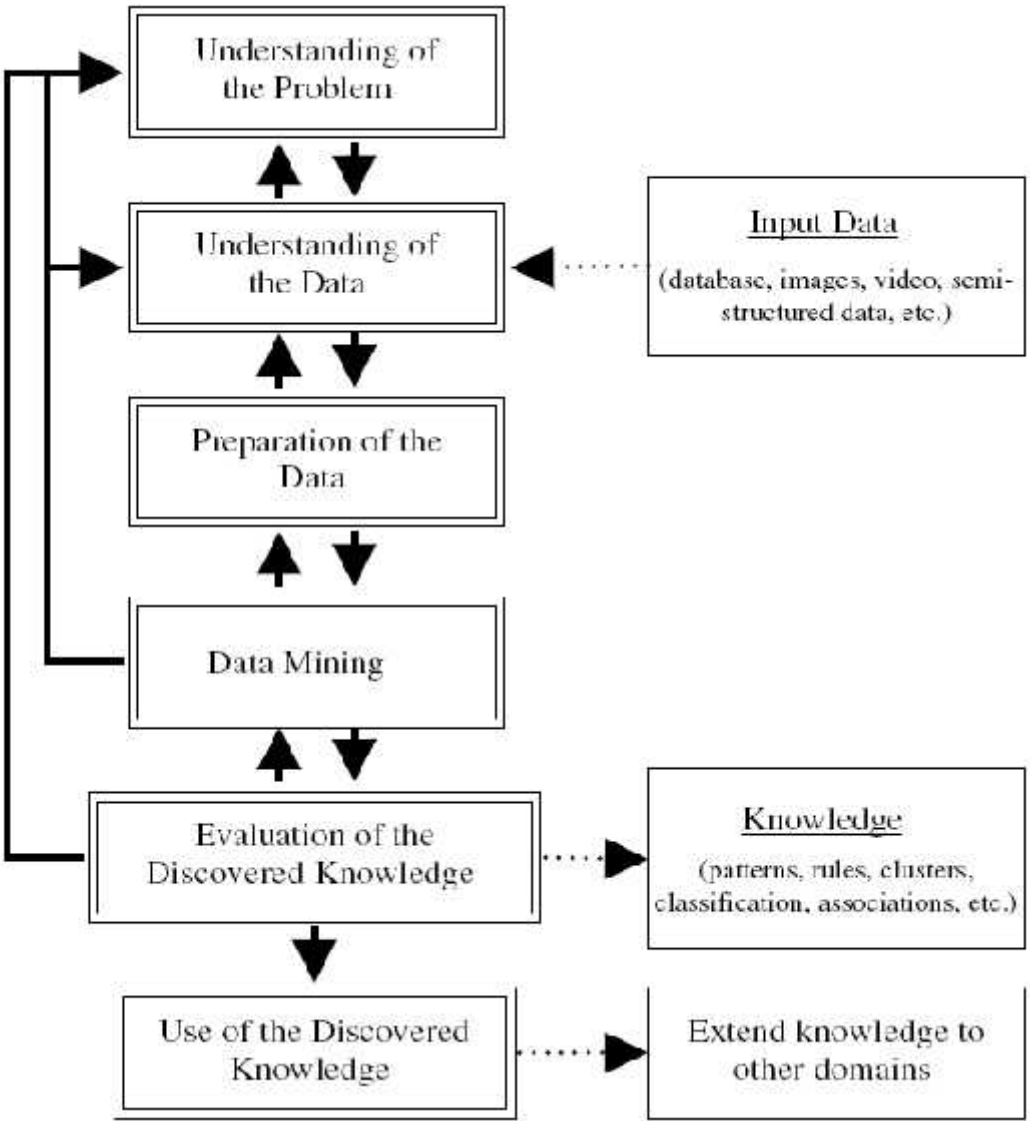
26. Anagaw S. Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project (BRHP).MA Thesis, Addis Ababa UniversityAddis Ababa, Ethiopia. June 2002.

27. Beshah T and Hill S. Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia. Unpublished Available on http://ai-d.org/pdfs/Beshah.pdf Accessed on April 2011

28. Tesso A. Application of Data Mining Technology To Identify Determinanat Risk Factors Of HIV Infection And To Find Their Association Rules: The Case of Center for Disease Control and Prevention (CDC).MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia. June 2005.

8. ANNEXES

Annex I –Hybrid Model Knowledge Discovery Process



The six-step Knowledge Discovery Process model. Source: Pal, N.R., Jain, L.C., (Eds.) 2005. Advanced Techniques in Knowledge Discovery and Data Mining, Springer Verlag.

Annex II

List of Domain Experts interviewed in the Data Mining (VCT) Process

- North Gondar Zone HIV/AIDS Prevention and Control Office (NGHAPCO)
- One Woreda HAPCO representative who are highest tested individual included(Gondar Town HAPCO)
- One NGO supporting VCT the center(SCN-E)
- One NGO providing VCT services in Gondar Town (FGA-E)
- One senior counselor(Gondar University Hospital, VCT Center)

Annex III

Questions for domain experts during understanding the domain step of the data mining process

1. How do you formulate HIV prevention and affected people care and support programme in your organization/offices? (i.e. what are the basis for these)?
2. Do you have reporting systems from VCT centers to support your planning and monitoring system?
3. If your answer is yes for question number 2, what are the issues you used for prevention and care and support planning from the VCT data from the hospital?
4. If your answer for question number 2 is yes, what are they?

Annex IV. Result from WEKA Explorer Window (Attributes/ Variables)

Attributes / Variables	Instances Full Descriptions	Abbreviated for data mining/ Category	Number of Instances	% of Instances
Sex of the client	Male	Male	6183	51.38
	Female	Female	5850	48.62
Residence	Gondar	Gondar	6955	57.80
	Dembia	Dembia	789	6.56
	Chilga	Chilga	686	5.70
	GondarZuria	Gzuria	641	5.33
	Lay Armachiho	LayArmachiho	625	5.19
	TacheArmachiho	TachAr	422	3.51
	Metema	Metema	357	2.96
	Wogera	Wogera	257	2.14
	East Belessa	Eastbelessa	216	1.80
	Takusa	Takusa	155	1.28
	Alefa	Alefa	120	1.00
	Dabate	Dabate	84	0.70
	Debark	Debark	70	0.58
	Tegede	Tegede	48	0.40
	Kuara	Kuara	43	0.36
	Janamora	Janamora	29	0.24
	Adrkaye	Adrkaye	25	0.21
	West Belessa	WBelessa	15	0.12
	East Armachiho	EastAr	8	0.07

	Beyeda	Beyeda	6	0.05
	Not mentioned /missed	Notmentioned	42	0.35
	South GondarWoredas	SGondar	182	1.51
	Other Amhara Region woredas	OtherAmhara	93	0.77
	Regions other than Amhara	Otherregion	154	1.28
	Other countries/abroad	Othercit	11	0.09
Religion	Orthodox	Orthodox	11337	94.22
	Muslim	Muslim	541	4.50
	Protestant	Protestant	72	0.60
	Jewish	Jewish	71	0.59
	Catholic	Catholic	12	0.10
Marital Status	Single	Single	6871	57.10
	Married	Married	2671	22.20
	Divorced	Divorced	1642	13.65
	Under 18 age	Underage	470	3.91
	Widowed	Widowed	320	2.66
	Not stated marital status/missed	Notcms	49	0.41
Educational Status	Secondary school	secondarys	3925	32.62
	Full Primary (grade 5-8)	FullP	2322	19.30
	Illiterate	Illiterat	1972	16.39
	Vocational (new 10 +1-4)	Vocanew	1038	8.63
	Tertiary level (12 +)	Tertiary	967	8.04
	First Primary (grade 1-4)	Firstp	957	7.95
	Kindergarten	KG	27	0.22

	Educational status not filled /missed	Checkes	825	6.86
Occupation/ Job	Student	Student	3252	27.03
	Farmer	Farmer	1709	14.20
	House wife	Housew	1564	13.0
	Daily Labourer	DayL	1087	9.03
	Business person	Business	933	7.75
	Dependent	Dependent	698	5.80
	Government	Government	486	4.04
	Teacher	Teacher	422	3.51
	Other not classified	Other	364	3.03
	Other Professional	Otherpro	216	1.8
	Armed (Police and military)	Armed	197	1.64
	Unskilled professional	Unskilled	189	1.57
	Driver	Driver	137	1.14
	Health professional	Healthpro	70	0.58
	Pension	Pension	39	0.32
	Not specified/missed	Notspe	670	5.57
History of STIs	No	No	11693	97.17
	Yes	Yes	340	2.83
Physical Status	Ok	Ok	11379	94.56
	Sick	Sick	654	5.44
Reason to test	To know their status	Tokn	10411	86.52
	Pre-marriage screening	Marriage	1467	12.19
	Partners pressure	Partner	66	0.55

	For ART	Art	40	0.33
	To confirm the previous test result	Toconfirm	25	0.21
	Revisit (unknown)	Revisit	11	0.09
	Reunion	Reunion	2	0.02
	Check reason of the client/missed	Checkres	12	0.01
Test Result	Negative	Negative	9668	80.35
	Positive	Positive	2093	17.39
	No result stated	Nores	272	2.26

Annex V. Rules from the J48 Classifier algorithm by using the whole attributes

Rule Number1 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to marriage then the result is negative (1445/54)

Rule Number2 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to Tokn (to know) then the result is negative (9525/1604)

Rule Number3 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to ART then the result is positive (27/3)

Rule Number4 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to partner then the result is negative (66/1)

Rule Number5 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to confirm and Maritalst (marital status) is equal to divorced then the result is negative (13/2)

Rule Number6 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to confirm and Maritalst (marital status) is equal to single then the result is positive (4/1)

Rule Number7 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to confirm and Maritalst (marital status) is equal to married then the result is negative (2/1)

Rule Number8 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to confirm and Maritalst (marital status) is equal to widowed then the result is positive (1)

Rule Number9 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to reunion then the result is positive (2/0)

Rule Number10 If the Pst (physical status of the clients) is equal to OK and Retotest (reason to test) is equal to revisit then the result is negative (9/2)

Rule Number11 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to yes then the result is positive (179/45)

Rule Number12 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 1-11 then the result is negative (15/7)

Rule Number13 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 11-21 then the result is negative (55/18)

Rule Number14 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 21-30 and sex is equal to Male then the result is negative (95/42)

Rule Number15 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 21-30 and sex is equal to Female then the result is positive (91/18)

Rule Number16 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 30-40 then the result is positive (110/34)

Rule Number17 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 40-50 then the result is positive (57/25)

Rule Number18 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 50-60 then the result is positive (17/6)

Rule Number19 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 60-70 then the result is positive (7/2)

Rule Number20 If the Pst (physical status of the clients) is equal to sick and STIHx (History of STIs) is equal to No and age 79-89 then the result is negative (1/0)

Annex VI. Age re-categorized for Apriori association algorithm

Age group	Category/for association	Number of instance	% of the instances
1-4	oneTofour	174	1.45
5-9	fiveTonine	227	1.89
10-14	tenTofourteen	184	1.53
15-19	fifteenTonineteen	2100	17.45
20-24	twentyTotwentyfour	3472	28.85
25-29	twentyfTotwentynine	2382	19.80
30-34	thirtyTothirtyfour	1242	10.32
35-39	thirtyfTothirtynine	936	7.78
40-44	fortyTofortyfour	561	4.66
45-49	fortyfTofortynine	355	2.95
50-54	fiftyTofiftyfour	184	1.53
55-59	fiftyfTofiftynine	87	0.72
60-64	sixtyTosixtyfour	72	0.60
65 and above	abovesixtyf	56	0.47

Annex VII. The best twenty Apriori association rules by using the whole dataset and selected 9 attributes

=== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: VCTTestassoc

Instances: 11761

Attributes: 9 (Age, Sex, Residence, Religon, Maritalst, Educst, Occupation, Retotest and Result

=== Associator model (full training set) ===

Minimum support: 0.45 and Minimum metric <confidence>: 0.9

Best rules found:

1. Sex=Male Result=Negative 5110 ==>Religon=Orthodox 4827 conf:(0.94)

2. Sex=Male Retotest=Tokn Result=Negative 4364 ==>Religon=Orthodox 4121 conf:(0.94)

3. Retotest=Tokn Result=Negative 8170 ==>Religon=Orthodox 7709 conf:(0.94)

4. Result=Negative 9668 ==>Religon=Orthodox 9118 conf:(0.94)

5. Sex=Male 6037 ==>Religon=Orthodox 5693 conf:(0.94)

6. Sex=Male Retotest=Tokn 5244 ==>Religon=Orthodox 4945 conf:(0.94)

7. Retotest=Tokn 10154 ==>Religon=Orthodox 9573 conf:(0.94)

8. Maritalst=Single Retotest=Tokn 5490 ==>Religon=Orthodox 5175 conf:(0.94)

9. Sex=Female Retotest=Tokn 4910 ==>Religon=Orthodox 4628 conf:(0.94)

10. Maritalst=Single Retotest=Tokn Result=Negative 4849 ==>Religon=Orthodox 4569 conf:(0.94)

11. Sex=Female 5724 ==>Religon=Orthodox 5390 conf:(0.94)

12. Sex=Female Result=Negative 4558 ==>Religon=Orthodox 4291 conf:(0.94)

13. Maritalst=Single 6736 ==>Religon=Orthodox 6338 conf:(0.94)

14. Maritalst=Single Result=Negative 6051 ==>Religon=Orthodox 5690 conf:(0.94)

15. Residence=GonderRetotest=Tokn Result=Negative 5006 ==>Religon=Orthodox 4648 conf:(0.93)

16. Residence=GonderRetotest=Tokn 6126 ==>Religon=Orthodox 5680 conf:(0.93)

17. Residence=Gonder Result=Negative 5614 ==>Religon=Orthodox 5191 conf:(0.92)

18. Residence=Gonder 6791 ==>Religon=Orthodox 6275 conf:(0.92)

19. Residence=GonderReligon=Orthodox 6275 ==>Retotest=Tokn 5680 conf:(0.91)

20. Residence=Gonder 6791 ==>Retotest=Tokn 6126 conf:(0.9)

Annex VIII. Best Rules from the Apriori association algorithm by usingAge, Sex, Marital status (Maritalst),Educational status (Educst), Occupation/Job and positiveResult attributes

=== Run information ===

Scheme: weka.associations.Apriori -N 50 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: VCTTestassoc

Instances: 2093

Attributes: 6(Age, Sex, Maritalst, Educst, Occupation, Result,)

=== Associator model (full training set) ===Apriori

Minimum support: 0.1 and Minimum metric <confidence>: 0.9

Best rules found:

1. Sex=Female 1166 ==> Result=Positive 1166 conf:(1)
2. Sex=Male 927 ==> Result=Positive 927 conf:(1)
3. Maritalst=Married 738 ==> Result=Positive 738 conf:(1)
4. Maritalst=Single 685 ==> Result=Positive 685 conf:(1)
5. Educst=Illiterat 610 ==> Result=Positive 610 conf:(1)
6. Educst=secondarys 527 ==> Result=Positive 527 conf:(1)
7. Age=twentyfTotwentynine 484 ==> Result=Positive 484 conf:(1)
8. Occupation=Housew 428 ==> Result=Positive 428 conf:(1)
9. Maritalst=Divorced 415 ==> Result=Positive 415 conf:(1)
10. Sex=Female Occupation=Housew 411 ==> Result=Positive 411 conf:(1)
11. Sex=Female Educst=Illiterat 377 ==> Result=Positive 377 conf:(1)
12. Educst=FullP 376 ==> Result=Positive 376 conf:(1)
13. Sex=Female Maritalst=Married 370 ==> Result=Positive 370 conf:(1)
14. Sex=Male Maritalst=Married 368 ==> Result=Positive 368 conf:(1)
15. Sex=Male Maritalst=Single 350 ==> Result=Positive 350 conf:(1)
16. Age=thirtyTothirtyfour 348 ==> Result=Positive 348 conf:(1)

17. Age=twentyTotwentyfour 338 ==> Result=Positive 338 conf:(1)

18. Sex=Female Maritalst=Single 335 ==> Result=Positive 335 conf:(1)

19. Age=twentyfTotwentynine Sex=Female 321 ==> Result=Positive 321 conf:(1)

20. Sex=Female Educst=secondarys 310 ==> Result=Positive 310 conf:(1)

21. Age=thirtyfTothirtynine 295 ==> Result=Positive 295 conf:(1)

22. Sex=Female Maritalst=Divorced 286 ==> Result=Positive 286 conf:(1)

23. Occupation=Business 284 ==> Result=Positive 284 conf:(1)

24. Occupation=Farmer 268 ==> Result=Positive 268 conf:(1)

25. Maritalst=Married Occupation=Housew 261 ==> Result=Positive 261 conf:(1)

26 Sex=Female Maritalst=Married Occupation=Housew 253 ==> Result=Positive 253 conf:(1)

27. Sex=Male Educst=Illiterat 233 ==> Result=Positive 233 conf:(1)

28. Age=twentyTotwentyfour Sex=Female 233 ==> Result=Positive 233 conf:(1)

29. Maritalst=Married Educst=Illiterat 232 ==> Result=Positive 232 conf:(1)

30. Occupation=DayL 230 ==> Result=Positive 230 conf:(1)

31. Sex=Male Educst=secondarys 217 ==> Result=Positive 217 conf:(1)

32. Sex=Male Occupation=Farmer 215 ==> Result=Positive 215 conf:(1)

33. Maritalst=Single Educst=secondarys 213 ==> Result=Positive 213 conf:(1)

34. Occupation=Student 211 ==> Result=Positive 211 conf:(1)

35. Maritalst=Married Occupation=Housew 261 ==> Sex=Female Result=Positive 253 conf:(0.97)

36. Maritalst=Married Occupation=Housew Result=Positive 261 ==> Sex=Female 253 conf:(0.97)

37. Maritalst=Married Occupation=Housew 261 ==> Sex=Female 253 conf:(0.97)

38. Occupation=Housew 428 ==> Sex=Female Result=Positive 411 conf:(0.96)

39. Occupation=Housew Result=Positive 428 ==> Sex=Female 411 conf:(0.96)

40. Occupation=Housew 428 ==> Sex=Female 411 conf:(0.96)

Annex IX. Best 20 Rules from the Apriori association algorithm by using 11 attributes and negative test result attributes

=== Run information ===

Scheme: weka.associations.Apriori -N 100 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: VCTTestassoc

Instances: 9668

Attributes: 11 (Age, Sex,Residence, Religion, Maritalst,Educst, Occupation/Job, STIHx,Pst,Retotestand negative Result)

=== Associator model (full training set) ===

Minimum support: 0.98 andMinimum metric <confidence>: 0.9

Best rules found:

1. STIHx=No 9543 ==> Result=Negative 9543 conf:(1)
2. Pst=Ok 9438 ==> Result=Negative 9438 conf:(1)
3. STIHx=No Pst=Ok 9358 ==> Result=Negative 9358 conf:(1)
4. Religon=Orthodox 9118 ==> Result=Negative 9118 conf:(1)
5. Religon=Orthodox STIHx=No 9004 ==> Result=Negative 9004 conf:(1)
6. Religon=Orthodox Pst=Ok 8899 ==> Result=Negative 8899 conf:(1)
7. Religon=Orthodox STIHx=No Pst=Ok 8827 ==> Result=Negative 8827 conf:(1)
8. Retotest=Tokn 8170 ==> Result=Negative 8170 conf:(1)
9. STIHx=No Retotest=Tokn 8049 ==> Result=Negative 8049 conf:(1)
10. Pst=Ok Retotest=Tokn 7949 ==> Result=Negative 7949 conf:(1)
11. STIHx=No Pst=Ok Retotest=Tokn 7870 ==> Result=Negative 7870 conf:(1)
12. Religon=Orthodox Retotest=Tokn 7709 ==> Result=Negative 7709 conf:(1)
13. Religon=Orthodox STIHx=No Retotest=Tokn 7598 ==> Result=Negative 7598 conf:(1)
14. Religon=Orthodox Pst=Ok Retotest=Tokn 7498 ==> Result=Negative 7498 conf:(1)
15. Religon=Orthodox STIHx=No Pst=Ok Retotest=Tokn 7427 ==> Result=Negative 7427 conf:(1)

16. Religon=Orthodox Pst=Ok 8899 ==>STIHx=No Result=Negative 8827 conf:(0.99)

17. Religon=Orthodox Pst=Ok Result=Negative 8899 ==>STIHx=No 8827 conf:(0.99)

18. Religon=Orthodox Pst=Ok 8899 ==>STIHx=No 8827 conf:(0.99)

19. Pst=Ok 9438 ==>STIHx=No Result=Negative 9358 conf:(0.99)

20. Pst=Ok Result=Negative 9438 ==>STIHx=No 9358 conf:(0.99)

Annex X. Best Rules from the Apriori association algorithm by using Age, Sex, Religion, Marital status (Maritalst), Educational status (Educst) Occupation/Joband Negative Result attributes

== Run information ===

Scheme: weka.associations.Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0

Relation: VCTTestassoc

Instances: 9668

Attributes: 7(Age, Sex, Religon, Maritalst, Educst, Occupation and Result (Negative))

=== Associator model (full training set) ===

Best 20 rules found:

1. Religon=Orthodox 9118 ==> Result=Negative 9118 conf:(1)

2. Maritalst=Single 6051 ==> Result=Negative 6051 conf:(1)

3. Religon=Orthodox Maritalst=Single 5690 ==> Result=Negative 5690 conf:(1)

4. Sex=Male 5110 ==> Result=Negative 5110 conf:(1)

5. Sex=Male Religon=Orthodox 4827 ==> Result=Negative 4827 conf:(1)

6. Sex=Female 4558 ==> Result=Negative 4558 conf:(1)

7. Sex=Female Religon=Orthodox 4291 ==> Result=Negative 4291 conf:(1)

8. Sex=Male Maritalst=Single 3301 ==> Result=Negative 3301 conf:(1)

9. Sex=Male Religon=Orthodox Maritalst=Single 3108 ==> Result=Negative 3108 conf:(1)

10. Age=twentyTotwentyfour 3067 ==> Result=Negative 3067 conf:(1)

11. Occupation=Student 2983 ==> Result=Negative 2983 conf:(1)

12. Sex=Male 5110 ==>Religon=Orthodox Result=Negative 4827 conf:(0.94)

13. Sex=Male Result=Negative 5110 ==>Religon=Orthodox 4827 conf:(0.94)

14. Sex=Male 5110 ==>Religon=Orthodox 4827 conf:(0.94)

15. Result=Negative 9668 ==>Religon=Orthodox 9118 conf:(0.94)

16. Sex=Male Maritalst=Single 3301 ==>Religon=Orthodox Result=Negative 3108 conf:(0.94)

17. Sex=Male Maritalst=Single Result=Negative 3301 ==>Religon=Orthodox 3108 conf:(0.94)

18. Sex=Male Maritalst=Single 3301 ==>Religon=Orthodox 3108 conf:(0.94)

19. Sex=Female 4558 ==>Religon=Orthodox Result=Negative 4291 conf:(0.94)

20. Sex=Female Result=Negative 4558 ==>Religon=Orthodox 4291 conf:(0.94)

Declaration

I, the undersigned, MPH in Health Informatics student declare that this thesis is my original work in partial fulfillment of the requirement for the degree of Master of Public Health in Health Informatics.

Name: Kinfe Wubetu

Signature: _____

Place of submission: Institute of Public Health, College of Medicine and Health Sciences, University of Gondar.

Date of Submission: _____

This thesis work has been submitted for examination with my/our approval as the University advisor (s).

Advisor(s):-

Name	**Signature**
Dr. Berihun Megabiaw (MD, MPH)	_____
Ato Atinkut Alamirrew (B.Sc., MPH-HI)	_____